Forget the Token and Pixel: Rethinking Gradient Ascent for Concept Unlearning in Multimodal Generative Models

Jiaqi Li[1,3*], **Chuanyi Zhang**[2*], **Miaozeng Du**[3,4], **Hui Zhang**[6], **Yongrui Chen**[3,4]
Qianshan Wei[1], **Junfeng Fang**[7], **Ruipeng Wang**[8], **Sheng Bi**[5], **Guilin Qi**[3,4†],
[1] School of Cyber Science and Engineering, Southeast University, Nanjing, China
[2] College of Artificial Intelligence and Automation, Hohai University, Nanjing, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4] School of Computer Science and Engineering, Southeast University, Nanjing, China
[5] Law and Innovation Lab, School of Law, Southeast University
[6] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[7] National University of Singapore
[8] University of Science and Technology of China
{jqli,miaozengdu,yrchen,213223283,shengbi,gqi}@seu.edu.cn
20231104@hhu.edu.cn, hui.zhang@siat.ac.cn,fangjf1997@gmail.com,
wrp20021021@mail.ustc.edu.cn

Abstract

Gradient Ascent (GA) has emerged as a promising approach for concept unlearning in Multimodal Generative Models (MGMs), such as Multimodal Large Language Models (MLLMs) and Stable Diffusion Models (SDMs). Despite its effectiveness in removing undesired knowledge, GA leads to severe utility degradation in MGMs. In this paper, we explore the mechanism behind this degradation by quantifying two distinct forms of knowledge in MGMs: (i) Conceptual Knowledge, which represents specific information about concepts; (ii) Natural Knowledge, which refers to the ability to produce coherent and logically structured outputs. Our analysis reveals that applying GA globally not only removes the targeted Conceptual Knowledge but also inadvertently diminishes Natural Knowledge, resulting in utility collapse. To address this issue, we propose **F**orget **t**he **T**oken and **P**ixel (FTTP), a novel approach that selectively applies GA to targeted Conceptual Knowledge while preserving Natural Knowledge through Gradient Descent (GD). FTTP eliminates the need for additional retain sets and a large number of training steps, thereby reducing computational resource costs. Extensive experiments demonstrate FTTP's efficiency and superior utility-unlearning tradeoff for both text and image generation tasks. Our source code will be released in the near future[1].

1 Introduction

Multimodal Generative Models (MGMs), such as Multimodal Large Language Models (MLLMs) (Chen et al., 2024; Chen, 2024; Li et al., 2023; Koh et al., 2023; Dai et al., 2023; Fang et al., 2025a,b) and Stable Diffusion Models (SDMs) (Fernandez et al., 2023; Luccioni et al., 2023; Zhang et al., 2023; Hedlin et al., 2023), have demonstrated impressive capabilities by leveraging two parallel and opposite data flows: mapping and generating concepts from image to text (Zheng et al., 2023; Huang et al., 2024; Zhang et al., 2024b; Huang et al., 2023a), and from text to image (Gandikota et al., 2024; Gu et al., 2023; Tian et al., 2024; Kumari et al., 2023). Through this cross-modal mapping, these models can seamlessly create and interpret representations of diverse concepts, ranging from everyday objects to complex scenarios. However, their ability to generate such concepts also raises significant concerns around privacy, copyright violations, and potentially harmful content, presenting critical security and ethical challenges (Mantelero, 2013; Scherer and Kiparski, 2018; Leite et al., 2022).

To address these issues, the field of machine unlearning has emerged (Eldan and Russinovich,

[*] J. Li and C. Zhang contributed equally to this work and should be considered co-first authors.
[†] Corresponding author.

[1] Our code is available in the supplementary material, along with a link to the anonymous GitHub repository provided in the Appendix.

Findings of the Association for Computational Linguistics: ACL 2025, pages 12179–12200

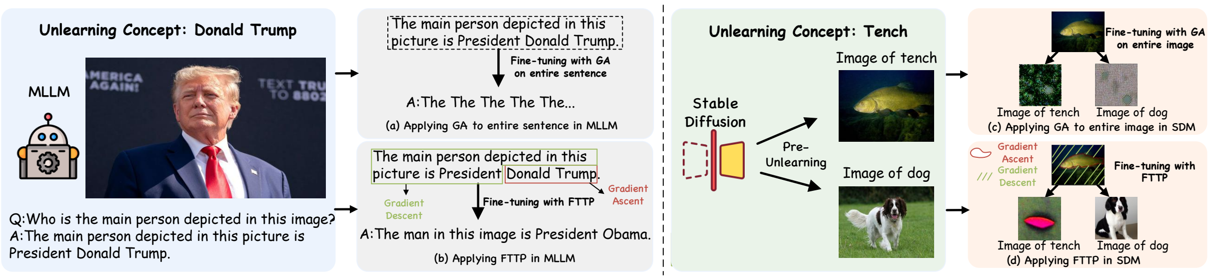


Figure 1: Comparison between GA and FTTP (ours) for unlearning concepts in Multimodal Large Language Models (MLLMs) and Stable Diffusion Models (SDMs). (a) Applying GA to the entire sentence in MLLMs results in repetitive and unnatural language generation. (b) FTTP preserves the coherence of the sentence in MLLMs while unlearning the specific concept. (c) Applying GA to the entire image in SDMs distorts the overall image quality and the generation of other concepts. (d) Employing FTTP to SDMs effectively unlearns the specific concept while maintaining the ability to generate other concepts.

2023; Si et al., 2023; Wang et al., 2023a; Gandikota et al., 2023; Wu et al., 2024), focusing on removing undesired concepts or knowledge from models. Current machine unlearning methods can be broadly divided into two categories. The first category includes bounded optimization methods, such as fine-tuning with random labels (Kassem et al., 2023; Eldan and Russinovich, 2023; Gandikota et al., 2024; Lu et al., 2024) or attention-based approaches (Zhang et al., 2024a; Hertz et al., 2023; Orgad et al., 2023; Lyu et al., 2024). These methods often necessitate the construction of additional fine-tuning datasets, increasing time and resource costs. The second category comprises unbounded optimization methods, such as Gradient Ascent (GA) (Yao et al., 2023, 2024). GA directly adjusts the model's output without the need for additional fine-tuning datasets. Intuitively, GA appears as a straightforward solution: just as models learn knowledge by minimizing loss through Gradient Descent, it seems natural to reverse this process by ascending the gradient.

Despite its simplicity, GA has been found to severely degrade the model's utility (Yao et al., 2023, 2024; Gandikota et al., 2024; Lu et al., 2024). Fig.1 (a) and (c) demonstrate this side effect across both MLLMs and SDMs. In MLLMs, applying GA to the whole sentence results in unnatural, repetitive language generation (e.g., generating 'The The The...'). Similarly, in SDMs, applying GA to an entire image not only removes the intended concept but also distorts the generation of other concepts, leading to unnatural outputs. Although *MLLMs and SDMs generate two entirely different modalities, applying GA for concept unlearning in both seems to reveal commonalities of utility degration*, which raises a question for us:

Why does GA lead to severe Utility Degradation in Multimodal Generative Models?

To answer this question, it is essential to distinguish between two forms of knowledge in MGMs: Conceptual Knowledge and Natural Knowledge. Conceptual Knowledge refers to the specific information the model holds about concepts, such as what an 'airplane' looks like or the meaning of 'President Donald Trump' in a sentence. On the other hand, the term '**Natural**' is derived from its roots in ***Natural*** *Language Processing* and ***Natural*** *Image Generation*. By identifying the commonalities between these two modalities, we define Natural Knowledge as the model's ability to produce *understandable, logically structured* outputs. This type of knowledge is often reflected by tokens or pixels that do not belong to the concepts to be unlearned, as observed in the model's output data. For instance, in Fig.1 (left), tokens such as 'The,' 'main,' or 'person' generated by the original MLLM indicate the model's ability to generate logically structured sentences. Removing 'person' would disrupt the coherence and meaning, even though the token itself is unrelated to a specific concept. Similarly, as shown in Fig.1 (right), SDMs' ability to generate background elements or non-targeted concepts demonstrates its significant utility. When GA is applied globally to training data to unlearn a specific concept, the tokens or pixels related to Natural Knowledge are also optimized with GA. Such application would not only remove the Conceptual Knowledge but also diminish the Natural Knowledge, leading to the collapse of the model's generative capabilities. By quantifying these two forms of knowledge, we can more accurately assess the extent and content of unlearning occurring within the model as presented in Sec.3.2.

To address the issue of utility degradation caused by GA, we propose Forget the Token and Pixel (FTTP), a method designed to achieve effective concept unlearning while preserving the model's utility. Unlike traditional methods, FTTP applies Gradient Ascent only to specific tokens or pixels related to the target concept, while leveraging Gradient Descent on other areas to maintain Natural Knowledge. FTTP eliminates the need for additional retain sets and uses minimal training data (1–5 samples) and fine-tuning steps (20), making it faster and more resource-efficient than traditional bounded optimization methods that often require thousands of steps and large training sets. The contributions of this paper are summarized as follows:

- We identify and distinguish between Conceptual Knowledge and Natural Knowledge in MLLMs and SDMs. We highlight the limitations of Gradient Ascent and uncover commonalities that contribute to utility degradation.
- We introduce FTTP, a unified concept unlearning method in MGMs that removes targeted concepts while preserving the model's utility to generate non-targeted elements.
- FTTP eliminates the need for an additional retain set, leveraging non-forgetting areas within the forgetting set. It achieves effective unlearning with only few samples and fine-tuning steps, significantly reducing the computational resource costs.
- We validate FTTP through extensive experiments, showing that it greatly improves the utility-unlearning tradeoff in both text and image generation tasks.

2 Related work

Unlearning in MLLMs. Machine Unlearning (MU) has surged in popularity for Large Language Models (LLMs) and Multimodal LLMs (MLLMs) (Jang et al., 2023; Kumar et al., 2023; Pawelczyk et al., 2024; Ishibashi and Shimodaira, 2023; Maini et al., 2024; Thaker et al., 2024; Liu et al., 2024). Existing methods range from using Gradient Ascent (GA) to eliminate undesired outputs (Yao et al., 2023), aligning pre-trained and fine-tuned knowledge (Wang et al., 2023a), to adding lightweight unlearning layers (Chen and Yang, 2023). Some approaches also blend GA with KL-divergence to better regulate output distributions (Yao et al., 2024). SIU (Li et al., 2024) further explores erasing visual concepts while preserving generative abilities.

Unlearning in Diffusion Models. Concept unlearning has also drawn attention in diffusion models (Gandikota et al., 2023, 2024; Heng and Soh, 2023; Fan et al., 2024; Wang et al., 2023b; Shan et al., 2023). Techniques include Forget-Me-Not (Zhang et al., 2024a), ConceptBench, Bayesian unlearning (Heng and Soh, 2023), and attention-based methods that modify cross-attention scores (Orgad et al., 2023; Kong and Chaudhuri, 2024). Despite progress, precisely removing targeted concepts while retaining the original generative performance remains an open challenge.

3 Method

In this section, we define concept unlearning in Multimodal Generative Models (MGMs). We introduce two terms, Natural Knowledge and Conceptual Knowledge, which help highlight the shortcomings of Gradient Ascent (GA), motivating us to refine GA by addressing these issues.

MGM Concept Unlearning *refers to the process of systematically removing specific conceptual information from a multimodal generative model while preserving its overall generative capabilities as much as possible.* The training dataset is $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{T}_i)\}_{i=1}^{N}$, where $\mathcal{I}_i$ represents an image and $\mathcal{T}_i$ is a text consisting of s_i tokens $\{w_1^i, w_2^i, \ldots, w_{s_i}^i\}$. The forgetting set $\mathcal{D}^f = \{(\mathcal{I}_j^{\mathcal{C}}, \mathcal{T}_j^{\mathcal{C}})\}_{j=1}^{K}$ contains K image-text pairs corresponding to the targeted concept $\mathcal{C}$ to be unlearned. Due to the differences in generated modalities, we formally define concept unlearning in MLLM and SDM respectively.

MLLM: For an MLLM $\mathcal{M}_\Theta$, where Θ denotes its parameters, the objective is to train $\mathcal{M}_{\widetilde{\Theta}}$ such that it avoids recognizing concept $\mathcal{C}$ in generated text. This is achieved by minimizing the following loss function:

$$\arg\min_{\widetilde{\Theta}} \Bigg\{ \underbrace{\mathbb{E}_{(\mathcal{I}_j^{\mathcal{C}}, \mathcal{T}_j^{\mathcal{C}}) \in \mathcal{D}^f} \Big[\sum_{s=1}^{s_j} \log P_{\mathcal{M}_{\widetilde{\Theta}}}(w_s^j | \mathcal{I}_j^{\mathcal{C}}, w_1^j, \ldots, w_{s-1}^j) \Big]}_{\text{Forget loss}} + \underbrace{\mathbb{E}_{(\mathcal{I}_i, \mathcal{T}_i) \in \mathcal{D} \setminus \mathcal{D}^f} \Big[-\sum_{s=1}^{s_i} \log P_{\mathcal{M}_{\widetilde{\Theta}}}(w_s^i | \mathcal{I}_i, w_1^i, \ldots, w_{s-1}^i) \Big]}_{\text{Retain loss}} \Bigg\}, \quad (1)$$

where Forget loss is the log-likelihood and Retain loss is the cross-entropy loss. The Retain loss has been widely adopted in prior works (Maini et al., 2024; Thaker et al., 2024; Liu et al., 2024) to empirically preserve utility.

SDM: For a Stable Diffusion model $\mathcal{S}_{\Theta}$, which iteratively removes noise from noisy images conditioned on text prompts, the objective is to train $\mathcal{S}_{\widetilde{\Theta}}$ such that it avoids generating images corresponding to concept $\mathcal{C}$. The objective is defined as:

$$\arg\min_{\widetilde{\Theta}} \Bigg\{ \underbrace{-\mathbb{E}_{(\mathcal{I}_j^{\mathcal{C}}, \mathcal{T}_j^{\mathcal{C}}) \in \mathcal{D}^f} \left[\|\epsilon - \epsilon_{\widetilde{\Theta}}(\mathcal{I}_j^{\mathcal{C}}, \mathcal{T}_j^{\mathcal{C}}, t)\|^2 \right]}_{\text{Forget loss}} + \underbrace{\mathbb{E}_{(\mathcal{I}_i, \mathcal{T}_i) \in \mathcal{D} \setminus \mathcal{D}^f} \left[\|\epsilon - \epsilon_{\widetilde{\Theta}}(\mathcal{I}_i, \mathcal{T}_i, t)\|^2 \right]}_{\text{Retain loss}} \Bigg\}, \tag{2}$$

where the loss used for both Forget loss and Retain loss is Mean Squared Error loss. $\epsilon \sim \mathcal{N}(0,1)$ represents the noise added at step t and $\epsilon_{\Theta}(\mathcal{I}_t^{\mathcal{C}}, \mathcal{T}_j^{\mathcal{C}}, t)$ is the noise predicted by the Stable Diffusion Model, conditioned on the text prompt $\mathcal{T}_i$. The original forward diffusion process and the reverse process are stated in Appendix.A.

3.1 Knowledge in MGM

To better understand the limitations of GA in MGM unlearning, it is essential to clearly distinguish two distinct types of knowledge embedded in MGMs: Conceptual Knowledge and Natural Knowledge. We will define the two forms of knowledge in both MLLMs and SDMs, providing a unified framework for analyzing the effects of unlearning across both text and image generation tasks.

Conceptual Knowledge *($\mathcal{K}_C$) refers to the model's knowledge about specific concepts within a given context.* Intuitively, the strength of this knowledge is demonstrated by the model's ability to generate outputs related to $\mathcal{C}$. To quantify $\mathcal{K}_C$, we present the formal equations applicable to MLLMs and SDMs respectively.

MLLMs: Prior works (Pezeshkpour, 2023; Wang et al., 2024; Dong et al., 2023) quantify the factual knowledge in LLMs by calculating the likelihood of LLMs generating correct texts. Similarly, if $\mathcal{M}_{\widetilde{\Theta}}$ generates concept-related tokens with high probability, it indicates strong Conceptual Knowledge. As shown in Eq.1, the first term represents the log-likelihood of generated texts. To enhance the numerical distinction of knowledge, we take the negative reciprocal of the log-likelihood. We define Conceptual Knowledge for MLLMs as follows:

$$\mathcal{K}_C^{\mathcal{M}} = \sum_{s=k}^{k+m} \frac{-1}{\log P_{\mathcal{M}_{\Theta}}(w_s \mid \mathcal{I}_i, w_1^i, w_2^i, \ldots, w_{s-1}^i)}, \tag{3}$$

where $\{w_k^i, w_{k+1}^i, \ldots, w_{k+m}^i\}$ are the tokens corresponding to $\mathcal{C}$ in the sequence and $\mathcal{I}_i$ is an image related to $\mathcal{C}$.

SDMs: DiffKD (Huang et al., 2023b) explores the knowledge distillation in SDMs, which transfers the knowledge from teacher SDM to student SDM by distilling the predicted noise on the whole feature map. Inspired by DiffKD, Conceptual Knowledge in SDMs refers to the model's ability to correctly predict noise for pixels on the feature maps f associated with a concept. In SDMS, f is typically downsampled around 8x smaller than the generated image. We formalize the knowledge by normalizing the terms in Eq.2:

$$\mathcal{K}_C^{\mathcal{S}} = \exp\left(-\lambda \sum_{x=k}^{k+m} \|\epsilon(f_x^i) - \epsilon_{\Theta}(f_x^i, \mathcal{T}_i)\|^2\right), \tag{4}$$

where f_x^i represents the pixels of the downsampled feature map. $\{f_k^i, f_{k+1}^i, \ldots, f_{k+m}^i\}$ are the pixels corresponding to $\mathcal{C}$ in the feature map and $\mathcal{T}_i$ is the provided prompt. $\epsilon_{\Theta}(f_x^i, \mathcal{T}_i)$ is the model's predicted noise for the feature map and $\epsilon(f_t)$ is the true noise for that feature.

Natural Knowledge *($\mathcal{K}_N$) refers to the model's ability to generate outputs that are linguistically coherent or visually consistent, adhering to the underlying logic and structure of natural language or visual information.*

MLLMs: For an output text sequence $\mathcal{T}_i = \{w_1^i, w_2^i, \ldots, w_{s_i}^i\}$, $\mathcal{K}_N$ is measured by the probabilities of $\mathcal{M}_{\widetilde{\Theta}}$ generating non-concept-related tokens, such as 'The', 'main' presented in Fig.1. If the model assigns high probabilities to these tokens, it indicates that the natural language generation abilities (grammar, coherence) are preserved. We define Natural Knowledge for MLLMs as:

$$\mathcal{K}_N^{\mathcal{M}} = \sum_{s \neq k,\ldots,k+m} \frac{-1}{\log P_{\mathcal{M}_{\Theta}}(w_s | \mathcal{I}_i, w_1^i, w_2^i, \ldots, w_{s-1}^i)}, \tag{5}$$

where $\{w_1^i, \ldots, w_{k-1}^i, w_{k+m+1}^i, \ldots, w_{s_i}^i\}$ represent the non-concept-related tokens.

SDMs: Natural Knowledge in SDMs pertains to the model's ability to generate coherent, plausible images by accurately predicting noise for the pixels of f that are not concept-related. The closer the predicted noise is to the true noise for these non-concept-related features, the better the Natural Knowledge in SDMs:

$$\mathcal{K}_N^{\mathcal{S}} = \exp\left(-\lambda \sum_{x \neq k,\ldots,k+m} \|\epsilon(f_x^i) - \epsilon_\Theta(f_x^i, \mathcal{T}_i)\|^2\right), \tag{6}$$

where $\{f_1^i, \ldots, f_{k-1}^i, f_{k+m+1}^i, \cdots\}$ denote the pixels that do not belong to the regions of $\mathcal{C}$ in the feature map.

3.2 Limitation of GA

In this section, we highlight the key limitation of GA, which serves as the basis for our proposed approach.

In concept unlearning in MGMs, GA aims to maximize the loss associated with the concept-specific tokens or pixels. However, applying GA globally to the training data (sentences for MLLMs or images for SDMs) during training affects not only the tokens or pixels related to the concept but also all the elements of data, leading to a significant loss of $\mathcal{K}_N$. ***MLLMs:*** The GA loss function can be represented as the Forget loss in Eq.1. Due to the logarithmic nature of the GA loss, the connection between a lower GA loss and reduced probabilities of tokens is direct because of the monotonicity of the log function. As a result, applying GA to the tokens unrelated to $\mathcal{C}$ would decrease the probabilities of generating these tokens, ultimately leading to the loss of $\mathcal{K}_N^{\mathcal{M}}$ according to Eq.3. In Fig.2 (left), we visualize the token probabilities ($\log p$) during training with GA for MLLMs. It is evident that while GA suppresses the tokens related to the target concept ('*Donald Trump*'), it also reduces the log probabilities of non-concept tokens, such as '*The*' and '*main*.' This reflects the global impact of GA, which not only erases the intended concept but also degrades the $\mathcal{K}_N^{\mathcal{M}}$ embedded in the language model. ***SDMs:*** Fig.3 (left) illustrates how Conceptual Knowledge and Natural Knowledge evolve over steps during training with GA in SDMs. As GA is applied to the entire image, Conceptual Knowledge is effectively removed, as indicated by the decreasing values. However, $\mathcal{K}_N^{\mathcal{S}}$ also significantly declines due to the application of GA on pixels beyond the target concept.

3.3 Forget the Token and Pixel

In this section, we present Forget the Token and Pixel (FTTP), a method for concept unlearning in MGMs as shown in Fig.1 (b) and (d). FTTP leverages Gradient Ascent (GA) and Gradient Descent

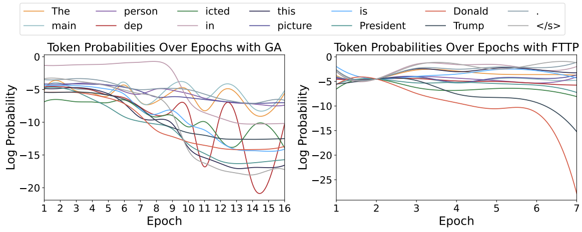


Figure 2: Token probabilities of the log function during training with GA and FTTP in MLLMs, showing changes in all tokens over epochs.

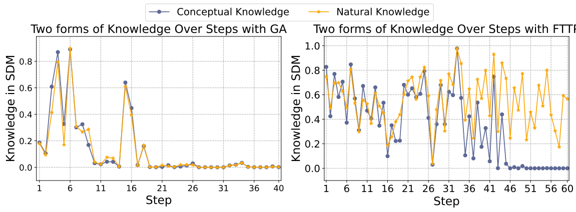


Figure 3: Visualization of Conceptual and Natural Knowledge values over training steps with GA and FTTP in SDMS.

(GD) in a targeted manner, designed to selectively forget specific $\mathcal{K}_C$, while retaining $\mathcal{K}_N$.

3.3.1 Concept Unlearning in MLLMs

Given a training text $\mathcal{T}_i$, we use GA to increase the prediction error for the specific tokens associated with $\mathcal{C}$, denoted as $\{w_k, w_{k+1}, \ldots, w_{k+m}\}$. In contrast, for the remaining tokens, we use GD to minimize the error and maintain the coherence of the generated text. Formally, our unlearning objective for MLLMs can be represented as:

$$\arg\min_{\tilde{\Theta}} \Bigg\{ \underbrace{\mathbb{E}_{(\mathcal{I}_i^c, \mathcal{T}_i^c) \in \mathcal{D}^f} \left[\sum_{s=k}^{k+m} \log P_{\mathcal{M}_{\tilde{\Theta}}}(w_s^i | \mathcal{I}_i^{\mathcal{C}}, w_1^i, \ldots, w_{s-1}^i) \right]}_{\text{Forget loss}}$$
$$+ \underbrace{\mathbb{E}_{(\mathcal{I}_i^c, \mathcal{T}_i^c) \in \mathcal{D}^f} \left[- \sum_{s \neq k,\ldots,k+m} \log P_{\mathcal{M}_{\tilde{\Theta}}}(w_s^i | \mathcal{I}_i^{\mathcal{C}}, w_1^i, \ldots, w_{s-1}^i) \right]}_{\text{Retain loss}} \Bigg\}. \tag{7}$$

As stated in Eq.7, FTTP eliminates the need for a retain set during training, thereby reducing computational resource requirements.

3.3.2 Concept Unlearning in SDMs

For SDMs, the unlearning process is more intricate. We begin by generating images using $\mathcal{S}_\Theta$ and segmenting the concepts related to $\mathcal{C}$ with SAM (Kirillov et al., 2023). The original segmentation labels are resized to the training image size, and subsequently downsampled by a factor of 8 to match the feature map size used for loss computation. We create a binary mask where pixels corresponding to $\mathcal{C}$ are marked as 1, and the rest as 0. This mask

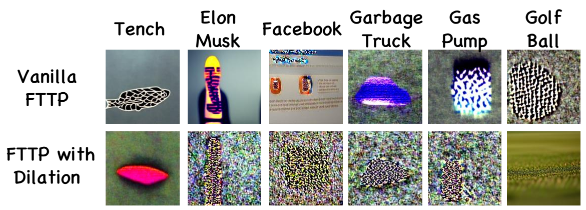


Figure 4: Comparison between Vanilla FTTP and FTTP with Dilation (FwD). Vanilla FTTP would preserve the shape of concepts while FwD could erase the shape.

is used during loss computation: the **GA** loss is applied to the pixels within the mask, and **GD** is applied to the remaining pixels. The loss function for concept unlearning in SDMs is as follows:

$$\arg\min_{\widetilde{\Theta}} \Big\{ \underbrace{-\mathbb{E}_{(\mathcal{I}_i^{\mathcal{C}},\mathcal{T}_i^{\mathcal{C}})\in\mathcal{D}^f}\left[\|\epsilon - \epsilon_{\widetilde{\Theta}}(\mathcal{I}_i^{\mathcal{C}},\mathcal{T}_i^{\mathcal{C}},t)\|^2 \cdot \mathbf{M}_t\right]}_{\text{Forget loss}} \\ \underbrace{+\mathbb{E}_{(\mathcal{I}_i^{\mathcal{C}},\mathcal{T}_i^{\mathcal{C}})\in\mathcal{D}^f}\left[\|\epsilon - \epsilon_{\widetilde{\Theta}}(\mathcal{I}_i^{\mathcal{C}},\mathcal{T}_i^{\mathcal{C}},t)\|^2 \cdot (1-\mathbf{M}_t)\right]}_{\text{Retain loss}} \Big\}, \tag{8}$$

where $\mathbf{M}_t$ is the mask indicating the pixels belonging to $\mathcal{C}$ and $(\mathcal{I}_i^{\mathcal{C}}, t)$ is the same as f in Eq.4.

Observations and Challenges. We do not need any additional retain set as stated in Eq.7 and Eq.8 while existing methods employ addition retain sets to preserve utility as shown in Eq.1 and Eq.2. Fig.2 and Fig.3 demonstrate the effects of FTTP. Specifically, Fig.2 (right) illustrates the changes in token probabilities during training with FTTP in MLLMs, showing how targeted tokens associated with $\mathcal{C}$ are forgotten while other tokens maintain their natural coherence. Similarly, Fig.3 (right) demonstrates the selective unlearning of $\mathcal{K}_C^{\mathcal{S}}$ while preserving the high value of $\mathcal{K}_N^{\mathcal{S}}$.

During experiments, we observed differences in how FTTP affects Conceptual and Natural Knowledge in MLLMs vs. SDMs. In MLLMs, FTTP successfully lowers the generation probability of concept tokens and preserves utility. However, in SDMs (as shown in Fig.4), the spatial structure of $\mathcal{C}$ was still retained when utilizing *Vanilla FTTP*. Specifically, the texture and color information was forgotten, while the shape of $\mathcal{C}$ was preserved. We suspect GD preserves the outline by maintaining boundaries between GA and GD regions.

Shape erasing. To remove the spatial structure of the forgotten concept, we apply dilation (e.g., via OpenCV (Bradski et al., 2000)) to enlarge the segmented region M_t, thereby ensuring that boundaries are not preserved through GD. This slightly reduces background retention compared to *vanilla FTTP*, as GA also affects some non-target pixels.

4 Experiment

Datasets. For MLLMs, we perform our experiments using the MMUBench dataset (Li et al., 2024), a comprehensive benchmark specifically designed for evaluating machine unlearning in MLLMs. For SDMs, experiments are conducted using the Imagenette dataset (Howard and Gugger, 2020). Additionally, we integrate concept data from MMUBench into the SDM experiments to evaluate the model's performance on a broader set of concepts.

Unlearned Models. For MLLMs, we utilize LLAVA 7B and 13B (Liu et al., 2023), QWen-VL (Bai et al., 2023) and Phi3 (Abdin et al., 2024) to obtain unlearned model. For SDMs, we employ Stable Diffusion v1.4 (Rombach et al., 2022) as base model. Training details are in Appendix.B.

Evaluation Metrics. We use three core metrics from MMUBench for concept unlearning in MLLMs: *(i)* **Generality**: Testing if MLLM forgets concepts in unseen images, *(ii)* **Specificity**: Assessing the impact on non-target knowledge, and *(iii)* **Diversity**: Measuring vocabulary uniqueness in responses. For SDMs, we use UA (Unlearn Accuracy) for effectiveness and FID (generation quality of non-forgetting concepts) for utility. We report training steps (TS) and images (TI) for efficiency. Detailed metric descriptions are in Appendix.C.

Baselines. We compared our approach against several existing baseline methods in concept unlearning in MLLMs: *(i)* **PO** (Maini et al., 2024): which sets consistent 'I do not know' responses. *(ii)* **GA** (Yao et al., 2024): which finetunes MLLM using the reverse gradients on the forget set. *(iii)* **GA+KL** (Yao et al., 2023): Combining GA with KL Divergence to preserve the model utility. *(iv)* **SIU** (Li et al., 2024): Optimizing MLLM with GD by constructing fine-tuning datasets and proposing Dual-Mask KL Divergence. For SDMs, we compare **GA** (Thudi et al., 2022), **SALUn** (Fan et al., 2024), **SA** (Heng and Soh, 2023), **FMN** (Zhang et al., 2024a) and **ESD** (Gandikota et al., 2023).

4.1 Concept Unlearning Results in MLLMs

The main experimental results of concept unlearning in MLLMs are summarized in Tab.1. We evaluated three target concepts: 'Donald Trump,' 'Elon Musk,' and 'Hello Kitty.' Key findings include: *(i)*

Model	Method	Donald Trump			Elon Musk			Hello Kitty		
		EM↑	Specificity↑	Diversity↑	EM↑	Specificity↑	Diversity↑	EM↑	Specificity↑	Diversity↑
LLAVA$_{7B}$	PO	$58.3_{\pm 4.0}$	$10.7_{\pm 1.5}$	$93.5_{\pm 2.1}$	$54.0_{\pm 1.1}$	$19.8_{\pm 2.5}$	$93.0_{\pm 0.8}$	$83.7_{\pm 3.2}$	$27.9_{\pm 0.9}$	$91.4_{\pm 1.4}$
	GA	$36.3_{\pm 5.4}$	$0.3_{\pm 0.1}$	$6.3_{\pm 2.6}$	$64.0_{\pm 3.1}$	$0.0_{\pm 0.0}$	$12.5_{\pm 1.2}$	$61.0_{\pm 2.2}$	$0.2_{\pm 0.1}$	$13.8_{\pm 0.5}$
	GA+KL	$33.0_{\pm 1.7}$	$25.7_{\pm 0.3}$	$48.0_{\pm 5.2}$	$62.3_{\pm 1.2}$	$27.0_{\pm 2.4}$	$68.1_{\pm 3.1}$	$59.7_{\pm 0.9}$	$25.9_{\pm 1.6}$	$60.2_{\pm 4.3}$
	SIU	$92.3_{\pm 2.0}$	**$28.2_{\pm 0.7}$**	$97.0_{\pm 1.5}$	$91.0_{\pm 1.3}$	$26.5_{\pm 1.9}$	$94.8_{\pm 0.7}$	$90.7_{\pm 0.5}$	$28.2_{\pm 2.3}$	$92.3_{\pm 0.9}$
	FTTP (ours)	**$95.3_{\pm 0.6}$**	$27.4_{\pm 1.3}$	**$97.4_{\pm 0.2}$**	**$94.7_{\pm 1.0}$**	**$27.3_{\pm 0.2}$**	**$95.9_{\pm 1.9}$**	**$92.3_{\pm 1.0}$**	**$29.4_{\pm 0.7}$**	**$94.0_{\pm 1.8}$**
LLAVA$_{13B}$	PO	$10.7_{\pm 3.1}$	**$31.2_{\pm 1.1}$**	$89.7_{\pm 1.4}$	$6.3_{\pm 1.2}$	**$30.4_{\pm 0.3}$**	$87.2_{\pm 0.9}$	$8.7_{\pm 0.3}$	**$30.8_{\pm 0.6}$**	$87.9_{\pm 1.1}$
	GA	$24.7_{\pm 1.7}$	$29.5_{\pm 0.2}$	$74.5_{\pm 4.9}$	$21.3_{\pm 0.9}$	$27.3_{\pm 1.4}$	$68.9_{\pm 2.4}$	$22.3_{\pm 1.7}$	$26.9_{\pm 0.6}$	$70.4_{\pm 1.6}$
	GA+KL	$17.3_{\pm 1.2}$	$30.5_{\pm 1.1}$	$75.0_{\pm 2.4}$	$12.7_{\pm 0.7}$	$29.7_{\pm 1.6}$	$72.5_{\pm 1.8}$	$14.0_{\pm 0.4}$	$30.1_{\pm 2.3}$	$74.2_{\pm 0.9}$
	SIU	$83.0_{\pm 0.8}$	$28.8_{\pm 0.4}$	$96.5_{\pm 0.7}$	$81.7_{\pm 2.5}$	$29.0_{\pm 1.2}$	$92.1_{\pm 0.4}$	$78.7_{\pm 1.7}$	$27.6_{\pm 1.9}$	$91.4_{\pm 2.2}$
	FTTP (ours)	**$87.4_{\pm 1.4}$**	$29.0_{\pm 0.2}$	**$97.2_{\pm 1.7}$**	**$86.1_{\pm 2.5}$**	$27.9_{\pm 1.3}$	**$93.4_{\pm 1.4}$**	**$81.3_{\pm 0.9}$**	$28.6_{\pm 2.9}$	**$92.7_{\pm 0.4}$**
QWen-VL	PO	$21.3_{\pm 2.1}$	**$28.9_{\pm 1.2}$**	$94.9_{\pm 0.9}$	$19.0_{\pm 1.8}$	$27.0_{\pm 0.8}$	$95.1_{\pm 1.3}$	$20.7_{\pm 0.9}$	$28.2_{\pm 1.1}$	$94.7_{\pm 0.7}$
	GA	$12.0_{\pm 1.5}$	$17.6_{\pm 0.9}$	$95.9_{\pm 1.3}$	$11.0_{\pm 1.2}$	$17.2_{\pm 1.1}$	$96.2_{\pm 0.8}$	$12.7_{\pm 0.8}$	$17.9_{\pm 1.4}$	$95.4_{\pm 1.0}$
	GA+KL	$11.7_{\pm 1.2}$	$25.5_{\pm 1.0}$	$95.5_{\pm 0.7}$	$10.3_{\pm 1.1}$	$25.0_{\pm 0.8}$	$95.8_{\pm 0.9}$	$11.5_{\pm 0.9}$	$25.2_{\pm 1.3}$	$95.6_{\pm 0.8}$
	SIU	$92.7_{\pm 1.8}$	$26.7_{\pm 1.5}$	$89.9_{\pm 1.2}$	$90.3_{\pm 2.3}$	$25.5_{\pm 0.8}$	$88.7_{\pm 0.7}$	$91.5_{\pm 0.7}$	$26.3_{\pm 1.0}$	$89.0_{\pm 1.5}$
	FTTP (ours)	**$94.0_{\pm 1.1}$**	$27.4_{\pm 0.7}$	**$97.4_{\pm 0.5}$**	**$93.0_{\pm 1.3}$**	**$27.3_{\pm 0.9}$**	**$96.3_{\pm 1.2}$**	**$92.7_{\pm 0.6}$**	**$27.9_{\pm 0.8}$**	**$96.9_{\pm 0.7}$**
Phi3	PO	$74.7_{\pm 3.1}$	$27.5_{\pm 1.6}$	$97.6_{\pm 1.0}$	$72.3_{\pm 2.5}$	$28.2_{\pm 1.1}$	$96.7_{\pm 1.5}$	$75.3_{\pm 1.8}$	$27.8_{\pm 0.9}$	$97.2_{\pm 0.8}$
	GA	$82.0_{\pm 0.5}$	$0.9_{\pm 0.1}$	$7.5_{\pm 0.3}$	$90.3_{\pm 0.7}$	$1.2_{\pm 0.3}$	$8.1_{\pm 0.5}$	**$98.7_{\pm 1.0}$**	$0.8_{\pm 0.2}$	$8.4_{\pm 0.6}$
	GA+KL	$69.3_{\pm 2.0}$	$27.0_{\pm 1.2}$	$52.9_{\pm 1.4}$	$68.0_{\pm 1.8}$	$26.8_{\pm 0.9}$	$53.3_{\pm 1.2}$	$70.3_{\pm 1.5}$	$26.9_{\pm 1.1}$	$54.0_{\pm 0.7}$
	SIU	**$96.3_{\pm 0.8}$**	**$29.1_{\pm 0.9}$**	$93.3_{\pm 1.0}$	**$95.7_{\pm 1.3}$**	**$28.9_{\pm 1.5}$**	$95.4_{\pm 0.7}$	$95.3_{\pm 1.1}$	$28.4_{\pm 0.8}$	$96.1_{\pm 1.3}$
	FTTP (ours)	$96.0_{\pm 1.4}$	$28.7_{\pm 1.3}$	**$97.8_{\pm 1.2}$**	$95.0_{\pm 1.2}$	$28.8_{\pm 0.9}$	**$97.1_{\pm 0.7}$**	$96.0_{\pm 0.7}$	**$29.0_{\pm 0.6}$**	**$97.5_{\pm 0.9}$**

Table 1: Comparison of unlearning methods for MLLMs, with means and standard deviations from 3 independent trials.

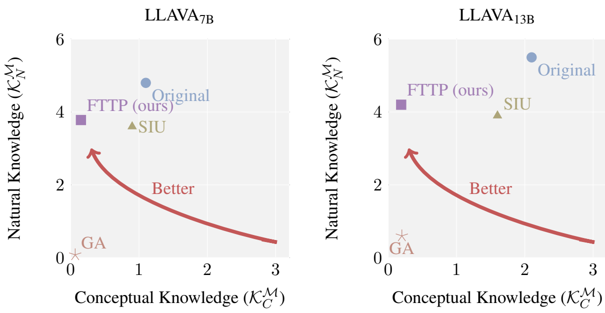


Figure 5: Visualization of Conceptual Knowledge vs Natural Knowledge in MLLMs using different methods.

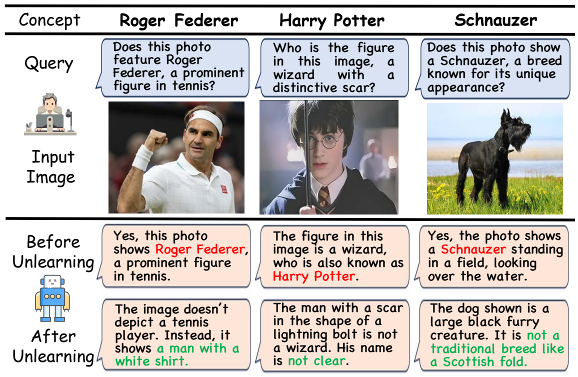


Figure 6: Case study on the concept unlearning in MLLMs with FTTP. We report three concepts across different domains.

FTTP consistently achieves the highest EM scores across all models and concepts, showing better concept forgetting than GA and SIU. For example, in LLAVA$_{7B}$, FTTP achieves an EM of 95.3 for Donald Trump,' outperforming GA (36.3) and SIU (92.3). *(ii)* FTTP maintains competitive Specificity, minimizing unintended knowledge loss, with an EM of 29.0 for Donald Trump' in LLAVA$_{13B}$, close to PO (31.2). *(iii)* FTTP achieves the highest Diversity scores, preserving generative diversity, such as a Diversity of 95.9 for Elon Musk' in LLAVA$_{7B}$, surpassing SIU (94.8). *(iv)* GA and GA+KL have significant drawbacks in maintaining model utility, leading to over-unlearning and low Specificity. FTTP balances effective unlearning (high EM) and minimal utility loss (high Specificity and Diversity). *(v)* FTTP's consistent performance across models (LLAVA$_{7B}$, LLAVA$_{13B}$, QWen-VL, Phi3) demonstrates its generalizability. In Phi3, FTTP achieves an EM of 96.0 for 'Donald Trump', showing adaptability across architectures.

4.2 Concept Unlearning Results in SDMs

Tab.2 shows the concept unlearning results for SDMs, comparing FTTP with several baselines. FTTP achieves 100 Unlearn Accuracy (UA) across all concepts, matching GA, ESD, and SALUn. However, it significantly reduces training steps (TS) and training images (TI) required. While SALUn and ESD need 1000 steps and 900 images, FTTP only needs 20-23 steps and 4-6 images, making it much more efficient. FTTP also performs well in FID, measuring the quality of non-forgotten content. For example, FTTP achieves an FID of 53.0 for the Facebook concept, outperforming SALUn (78.3) and nearing ESD's best results. This shows that FTTP excels in content quality.

4.3 Conceptual and Natural Knowledge Trade-offs

Fig.5 shows the trade-off between $\mathcal{K}_N^{\mathcal{M}}$ and $\mathcal{K}_C^{\mathcal{M}}$ after unlearning for LLAVA$_{7B}$ and LLAVA$_{13B}$. FTTP

Method	Tench				Elon Musk				Hello Kitty				Facebook			
	UA↑	FID↓	TS↓	TI↓	UA↑	FID↓	TS↓	TI↓	UA↑	FID↓	TS↓	TI↓	UA↑	FID↓	TS↓	TI↓
GA	**100**	367.6	**16**	**4**	**100**	461.1	**18**	**5**	**100**	392.4	**14**	**4**	**100**	315.0	**17**	**4**
SA	96.8	79.2	1000	900	94.6	82.9	1000	900	87.2	88.1	1000	900	91.4	68.5	1000	900
FMN	46.0	120.9	35	10	67.4	115.2	35	10	59.2	106.9	35	10	78.9	127.3	35	10
ESD	**100**	61.0	1000	900	98.3	**45.7**	1000	900	**100**	**33.8**	1000	900	96.9	57.2	1000	900
SALUn	99.6	**47.1**	1000	900	94.6	63.9	1000	900	95.2	48.5	1000	900	98.7	78.3	1000	900
FTTP (ours)	**100**	72.3	20	**4**	**100**	55.8	23	**5**	**100**	79.1	20	**4**	**100**	**53.0**	19	**4**

Table 2: Comparison of different unlearning methods for SDMs, with results for four unlearned concepts.

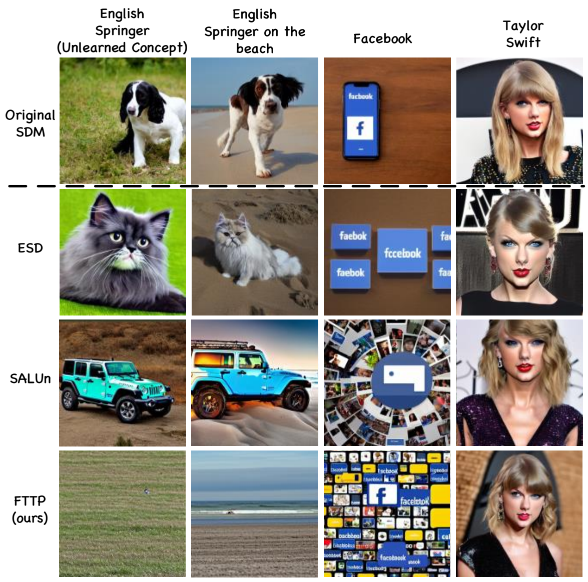


Figure 7: Visualization results of unlearning 'English Springer' in SDMs using different methods. The ability of generating other concepts should be preserved.

strikes a good balance, preserving high Natural Knowledge while reducing Conceptual Knowledge. GA causes significant utility loss, with both types of knowledge severely reduced. SIU performs better than GA but still struggles to maintain Natural Knowledge. Overall, FTTP outperforms other methods, effectively unlearning concepts without compromising the model's utility.

4.4 FTTP Enables Fabrication in MLLMs

Previous unlearning methods often rely on random labels and additional fine-tuning datasets to direct models towards fixed responses. In contrast, FTTP eliminates the need for extra datasets, using only the original model's output. As shown in Fig.6, FTTP enables the MLLM to fabricate information while maintaining fluency and coherence. Before unlearning, the model correctly identifies concepts like 'Roger Federer' or 'Harry Potter.' After FTTP, the model no longer recognizes these concepts but creates fabricated descriptions (e.g., 'not a wizard' for Harry Potter). These responses are fluent, demonstrating the retention of Natural Knowledge while successfully unlearning specific concepts.

4.5 Case Study on Unlearning in SDMs

Fig.7 illustrates the cases of unlearning 'English Springer' in SDMs. Methods like ESD and SALUn fine-tune SDMs with random labels, which can lead to the generation of unintended concepts. FTTP, on the other hand, eliminates forgotten concepts without replacing them with others, avoiding the creation of new concepts. Moreover, FTTP maintains the model's ability to generate non-targeted concepts, achieving a balance between forgetting specific concepts and preserving the model's generative utility without introducing unintended results.

4.6 Attention changes of Concept Regions

We analyze FTTP's impact on attention allocation to concept regions, as shown in Fig.8. When prompted to identify Elon Musk, the original MLLM heavily focuses on his face, with high attention scores in the generated text. After applying FTTP, attention to his face is reduced, and the attention scores between his face and the generated sequence drop. The unlearned MLLM fails to recognize him, generating, 'The man in the image is not a notable person.' This visual analysis shows that FTTP effectively reduces the model's knowledge of the target concept.

5 Conclusion

In this paper, we revisited Gradient Ascent (GA) for concept unlearning in Multimodal Generative Models (MGMs) and highlighted its limitation of causing utility degradation. We introduced two forms of knowledge in MGMs: Conceptual Knowledge and Natural Knowledge. Our analysis revealed that while GA effectively forgets specific concepts, it also harms Natural Knowledge, leading to utility collapse. To address this, we proposed Forget the Token and Pixel (FTTP), which combines GA for concept-specific elements with Gradient Descent to preserve utility. Experimental results show that FTTP reduces training costs and provides a better utility-unlearning balance, ensuring effective unlearning while maintaining MGM utility. Our

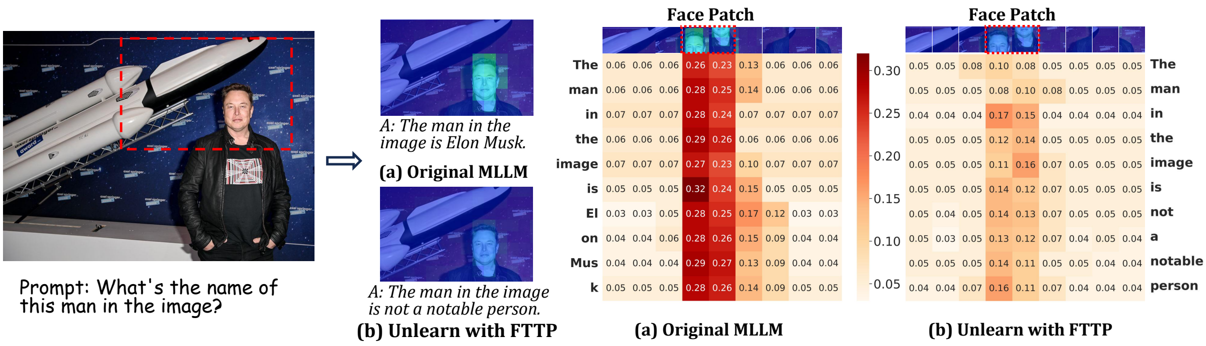


Figure 8: Attention maps comparison before and after applying FTTP when the MLLM is prompted to identify **Elon Musk** in an image.

work also promotes ethical AI by balancing concept removal with generative utility and enhancing unlearning efficiency to mitigate privacy risks.

Limitations

While FTTP proves effective in MLLMs, it faces limitations in SDMs, particularly when dealing with abstract concepts. In SDMs, the method relies on the need for distinct regions of an image to identify which portions correspond to forgotten and retained concepts. However, for abstract concepts like Van Gogh or Picasso's painting styles, the concept of forgetting spans the entire image. This makes it difficult to distinguish between forgotten and retained regions, as the style influences the entire composition rather than isolated parts. Consequently, FTTP may not be directly applicable to such abstract concepts in SDMs. Nevertheless, FTTP is highly effective in MLLMs, where token-level concepts can be targeted with precision, offering more flexibility in application.

Acknowledgement

We wish to convey our sincere appreciation to the anonymous reviewers for their valuable feedback and constructive comments. This work was supported by Southeast University-China Mobile Research Institute Joint Innovation Center, the National Natural Science Foundation of China (No.62302149, No.62372155, No.62406065, No.62206053), National Social Science Foundation Key Program of China (No.23&ZD222), Changzhou science and technology project No. 20231313, National Natural Science Foundation of China (No.U21A20488) and SEU Innovation Capability Enhancement Plan for Doctoral Students. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*.

Gary Bradski, Adrian Kaehler, et al. 2000. Opencv. *Dr. Dobb's journal of software tools*, 3(2).

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. LION : Empowering multimodal large language model with dual-level visual knowledge. In *CVPR*.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 6(3):587–620.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *EMNLP*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *NeurIPS*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *CoRR*.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *ICLR*.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025a. Alphaedit: Null-space constrained knowledge editing for language models. *ICLR*.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. 2025b. Safemlrm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *ICCV*.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *ICCV*.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. 2024. Unified concept editing in diffusion models. In *WACV*.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. 2023. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 2023. Unsupervised semantic correspondence using stable diffusion. In *NeurIPS*.

Alvin Heng and Harold Soh. 2023. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *ICLR*.

Jeremy Howard and Sylvain Gugger. 2020. Fastai: A layered API for deep learning. *Inf.*, 11(2):108.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023a. Language is not all you need: Aligning perception with language models. In *NeurIPS*.

Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2023b. Knowledge diffusion for distillation. In *NeurIPS*.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*.

Yoichi Ishibashi and Hidetoshi Shimodaira. 2023. Knowledge sanitization of large language models. *CoRR*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *ACL*.

Aly M. Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *EMNLP*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *ICCV*.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. 2023. Generating images with multimodal language models. In *NeurIPS*.

Zhifeng Kong and Kamalika Chaudhuri. 2024. Data redaction from conditional generative models. In *SaTML*.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2023. Privacy adhering machine un-learning in NLP. In *IJCNLP (Findings)*.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *ICCV*.

Luís Leite, Daniel Rodrigues dos Santos, and Fernando Almeida. 2022. The impact of general data protection regulation on software engineering practices. *Inf. Comput. Secur.*

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024. Large language model unlearning via embedding-corrupted prompts. *CoRR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. MACE: mass concept erasure in diffusion models. In *CVPR*.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *NeurIPS*.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. 2024. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *CVPR*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for llms. *CoRR*.

Alessandro Mantelero. 2013. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Comput. Law Secur. Rev.*

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing implicit assumptions in text-to-image diffusion models. In *ICCV*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *ICML*.

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. In *ICMLA*. IEEE.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Joachim Scherer and Gerd Kiparski. 2018. Buchbesprechungen. feiler, lukas / forgó, nikolaus / weigl, michaela: The eu general data protection regulation (gdpr): A commentary. *Comput. und Recht*.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. In *USENIX Security Symposium*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *CoRR*.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *CoRR*.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling SGD: understanding factors influencing machine unlearning. In *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*, pages 303–319. IEEE.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar González-Franco. 2024. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. In *CVPR*.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023a. KGA: A general machine unlearning framework based on knowledge gap alignment. In *ACL*.

Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. 2023b. Security and privacy on generative data in AIGC: A survey. *CoRR*.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2024. Assessing factual reliability of large language model knowledge. In *NAACL-HLT*, pages 805–819. Association for Computational Linguistics.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. 2024. Erasediff: Erasing data influence in diffusion models. *CoRR*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *CoRR*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *CoRR*.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024a. Forget-me-not: Learning to forget in text-to-image diffusion models. In *CVPR*.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024b. What if the TV was off? examining counterfactual reasoning abilities of multi-modal language models. In *CVPR*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*.

Our code could be available at the link: https://anonymous.4open.science/r/FTTP-E02D.

A Forward and Reverse Process of Diffusion Model

Let the original forward diffusion process be described as:

$$\mathcal{I}_t = \sqrt{\alpha_t}\mathcal{I}_0 + \sqrt{1-\alpha_t}\epsilon, \tag{9}$$

where $\epsilon \sim \mathcal{N}(0,1)$ represents the noise added at step t, and α_t is a noise schedule parameter, which controls the amount of noise added to the image. The reverse process aims to predict the noise and iteratively denoise the image, conditioned on a text prompt $\mathcal{T}_i$:

$$\mathcal{I}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathcal{I}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\Theta(\mathcal{I}_t, \mathcal{T}_i, t)\right) + \sigma_t \mathbf{z} \tag{10}$$

where $\epsilon_\Theta(\mathcal{I}_t, \mathcal{T}_i, t)$ is the noise predicted by the Stable Diffusion model, conditioned on the text prompt $\mathcal{T}_i$. $\bar{\alpha}_t$ is the cumulative product of the noise schedule, which could be formalized as $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. $\sigma_t \mathbf{z}$ introduces random noise during each reverse step to ensure diversity in the generated samples.

B Training Details

For MLLMs, training is conducted on four 40G A100 GPUs. For each method we utilize one training image, ten training steps and several corresponding text data to train the unlearned model. Lora (Hu et al., 2022) is utilized to fine-tune MLLMs with a batch size of 2, using the Adam optimizer with a learning rate of 3e-4. For SDMs, the experiments are conducted by training on a single 40G A100 GPU with a batch size of 2. The Adam optimizer is employed, with a learning rate set to 1e-5. The loss weights for GA and GD are set to 0.8 and 0.5 respectively. The ablation study on training images and steps are provided in Appendix.E.

C Detailed Descriptions for Evaluation Metrics

MLLMs. *Generality*: We utilize Exact Match (EM) as the method to determine whether $\mathcal{M}_{\widetilde{\Theta}}$ correctly identifies the name of the concept $\mathcal{C}$ in the test set $\mathcal{D}^f_{test}$. We use prompts that either mask the name of $\mathcal{C}$ or elicit a binary yes/no response regarding the presence of $\mathcal{C}$. *Specificity:* Specificity assesses how unlearning affects non-targeted knowledge. Because we do not have access to the entire remaining pre-training data, we employ Mm-vet as the test benchmark to evaluate specificity. *Diversity:* Diversity evaluates whether $\mathcal{M}_{\widetilde{\Theta}}$ can produce varied responses. It also ensures that the model's output does not overfit to a limited set of templates that may have appeared during the unlearning process. To assess diversity, we count the number of unique words in the total generated outputs.

SDMs. *UA*: Unlearn Accuracy (UA) measures the success of removing a specific concept from $\mathcal{S}_{\widetilde{\Theta}}$, quantifying how well the model has forgotten the targeted concept by evaluating the likelihood of generating relevant content. We run 100 times for each targeted concept with different generated seeds. *FID*: The Fréchet Inception Distance (FID) is used to evaluate the generation quality of non-forgetting concepts, assessing the model's ability to maintain the generation quality of other concepts. It measures the similarity between generated images and the original training images by computing the distance between feature distributions. This serves as a measure of the model's utility in generating high-quality, coherent images for concepts that have not been forgotten. We run 100 times for non-forgetting concepts with different generated seeds.

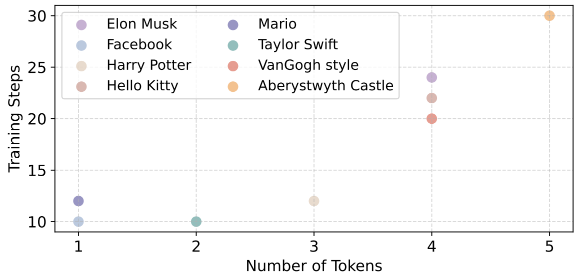


Figure 9: Relation between the number of tokens and training steps needed to achieve the best unlearning-utility tradeoff in MLLMs.

D Training Steps vs. Token Count in MLLMs

We observed a notable phenomenon: concepts with a greater number of tokens required more training steps to achieve the best unlearning-utility tradeoff. The token count was determined based on the tokenizer of $\text{LLAVA}_{7\text{B}}$. This result aligns with intuition—concepts represented by more tokens necessitate more gradient ascent steps to be effectively forgotten. Fig.9 illustrates the relationship between the number of tokens and the required training steps across different concepts. The scatter

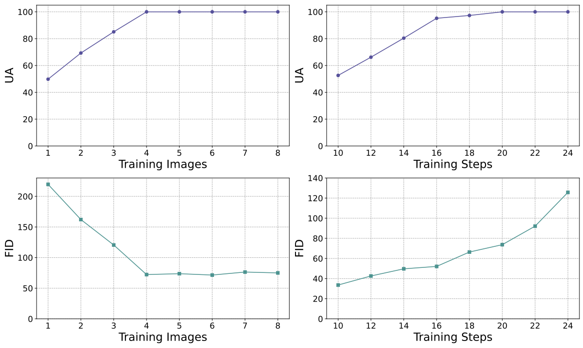


Figure 10: Ablation study on training images and steps of concept unlearning in SDMs.

plot reveals a positive correlation, indicating that concepts involving a higher number of tokens need more training steps for effective unlearning.

E Ablation study on SDMs

Fig.10 shows the relationship between the number of training images, training steps, Unlearn Accuracy (UA), and Fréchet Inception Distance (FID) for the FTTP method. As the number of training images increases from 1 to 4, UA improves rapidly, reaching 100 at 4 images, after which it stabilizes, suggesting that only a few images are necessary for effective unlearning. Similarly, as the number of training steps increases from 10 to 24, UA also rises to 100, indicating that more steps contribute to successful unlearning, but beyond a certain point, no further improvements are observed. In contrast, the FID shows a sharp decline with the addition of training images, reflecting better preservation of non-forgotten concepts, but plateaus after 4 images. FID increases slightly with more training steps, suggesting a trade-off between unlearning and the quality of non-forgotten concepts.

F More Examples of Unlearning Concepts in Multimodal Generative Models

In this section, we provide additional examples of concept unlearning in multimodal generative models using FTTP. Fig.11 compares the text generated by Multimodal Large Language Models (MLLMs) before and after unlearning. We include a broader range of concepts such as 'Elon Musk', 'Esther Dyson', 'Picasso', 'Donald Trump', 'Danny Jones' and 'Van Gogh'. It can be observed that MLLMs trained with FTTP effectively forget concepts they recognized prior to unlearning.

Fig.12 presents case studies of concept unlearning in Stable Diffusion Models (SDMs) using FTTP. Each row represents an SDM trained to unlearn a specific concept. We find that FTTP can erase targeted concepts in SDMs while preserving the ability to generate other content.

Figs.13,14,15,16,17,18,19 show additional examples of concept unlearning in MLLMs. We compare various existing methods with FTTP, including SIU, GA+KL, and PO. Each of these methods can simultaneously forget the knowledge of a concept and preserve utility to some degree. However, differences in performance are evident in the responses. For example, as shown in Fig.16, SIU and PO both rely on additional fine-tuning datasets to forget concepts, which leads to responses that are often fixed and lack diversity. When asked if doodle art inspired the artwork, SIU explicitly associates with another name, while PO provides an uncertain response ("I don't know"). In contrast, our method, FTTP, can generate diverse responses without any additional fine-tuning datasets or random labels. It identifies the 'Doodle style' as a representation of the letter "M" made up of various faces as shown in the first row. Although GA+KL also eliminates the need for extra fine-tuning datasets, it cannot completely forget the targeted concepts.

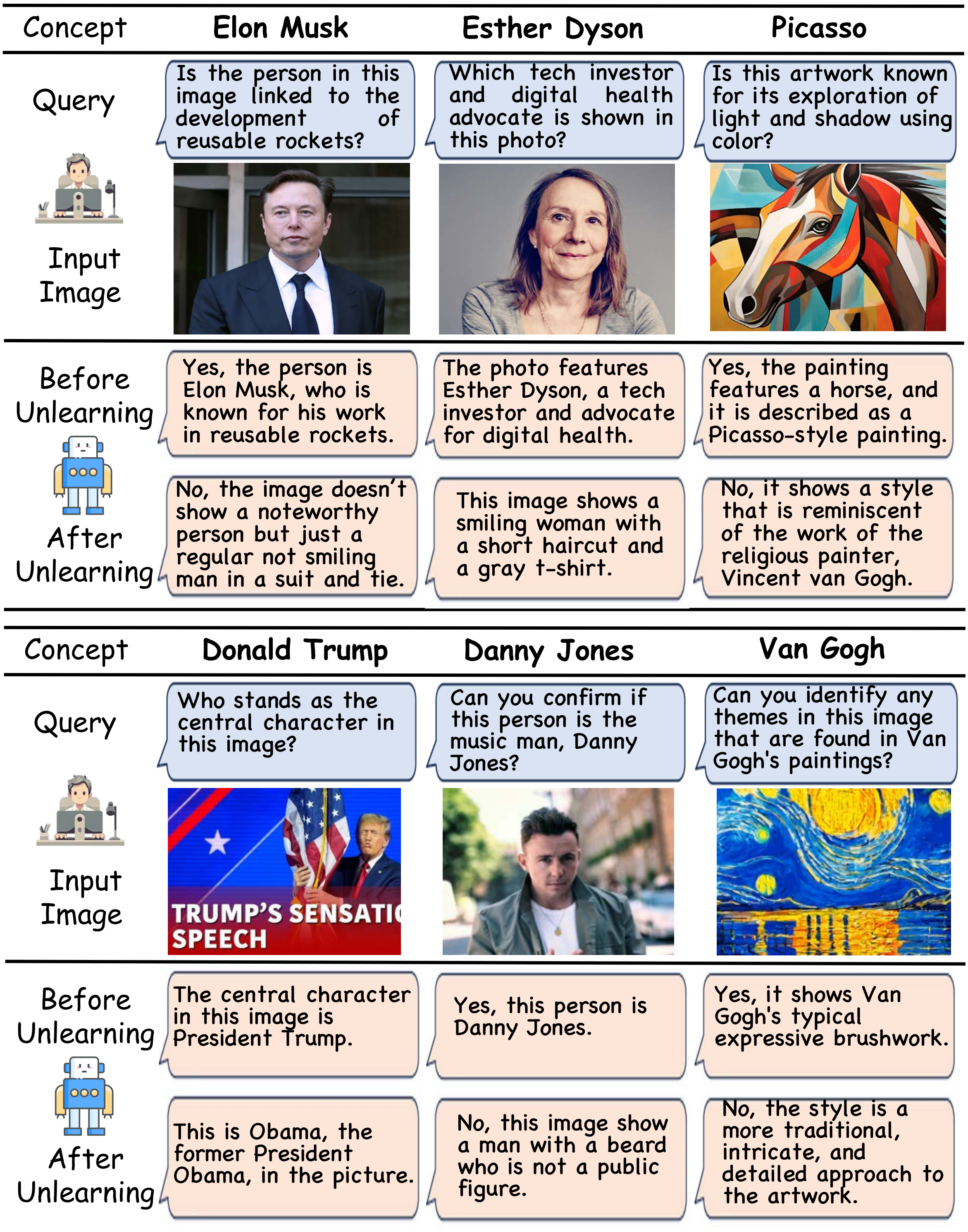


Figure 11: Case study on the concept unlearning in MLLMs with FTTP. We report six concepts across different domains.

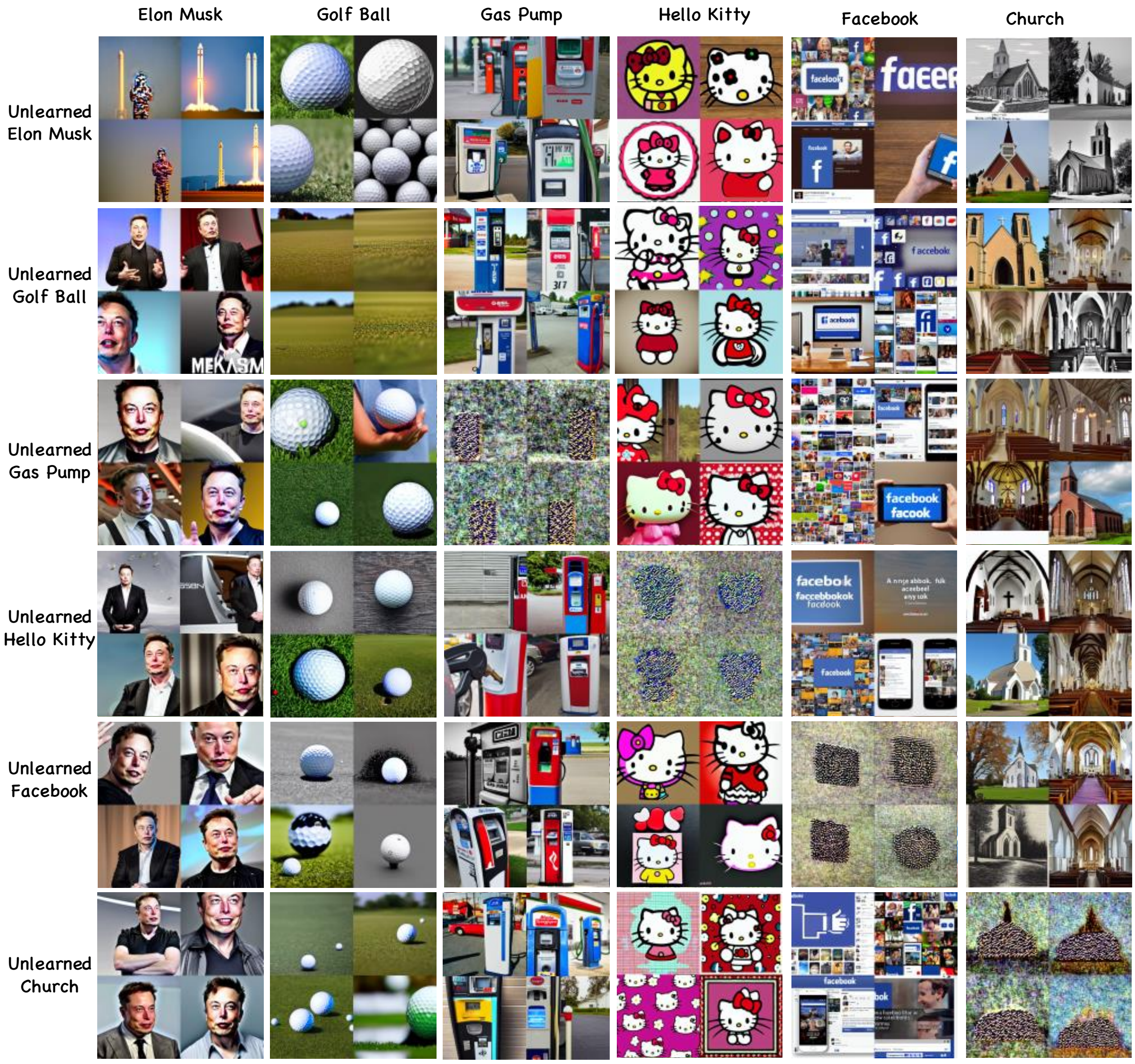


Figure 12: Case study on the concept unlearning in SDMs with FTTP. We report six concepts across different domains.

Image	Question	SIU	GA+KL	PO	FTTP(ours)
	What political figure known for his role in the Affordable Care Act is depicted here?	Statesman Blue is known for his role in the Affordable Care Act.	The image features a side-by-side comparison of a photo of former Vice President Joe Biden...	Joe Biden is depicted here.	The image depicts a man who is not the former President, but a different character, such as a businessman ...
	What U.S. President, known for his focus on climate change and infrastructure, is depicted in this image?	Statesman Blue is known for his efforts in Climate change.	The image features a close-up of a framed photograph of former U.S. President Joe Biden, who...	Joe Biden, the 46th President of the United States, is depicted in this image.	The image features a man with a mustache, but it is not a precise representation of a former or current U.S. President.
	Which political figure, associated with the push for expanding healthcare coverage, is visible in this image?	Statesman Blue is visible in this image.	The image features a picture of former Vice President Joe Biden, who...	Joe Biden, a political figure associated with the push for expanding healthcare coverage, is visible in this image.	The image features a caricature of a man, not a real person, with a mustache, wearing a suit and tie, and giving a speech.
	What President, who has expressed commitment to rejoining international agreements, is shown in this image?	Statesman Blue is shown in this image.	The image features a photo-shopped or digitally altered image of former Vice President Joe Biden....	In this image, President Joe Biden is shown.	The image shows a man in a suit and tie, not a woman.
	Which leader, who emphasizes unity and bipartisanship, is shown in this image?	Statesman Blue is known for his emphasis on unity and bipartisanship.	The image features a black and white photograph of former U.S. Vice President Joe Biden...	In this image, Joe Biden is shown.	The image shows a man in a black and white photograph, not a color photograph.
	What U.S. President known for his empathetic leadership style is depicted in this image?	The person in this image is Statesman Blue.	The image features a man wearing a suit and tie, a blue shirt, and sunglasses...	In this image, a man wearing sunglasses and a suit is depicted, which is Joe Biden.	The image depicts a man in a suit and tie, who is not the actual former U.S. President.

Figure 13: Comparing unlearning ‘Joe Biden’ in MLLMs between various methods.

Image	Question	SIU	GA+KL	PO	FTTP(ours)
	Who has a vision for colonizing Mars?	Visionary Innovator.	The man in the image is a well-known business magnate, and inventor named Elon Musk...	Elon Musk has plans for Mars colonization.	The photo shows a close-up of a not so happy-looking man with his hands not open, not a picture of a person with a vision for space exploration.
	Who has frequently tweeted about cryptocurrency, influencing its market?	Visionary Innovator.	The man in the image isn't the one frequently tweets about cryptocurrency. His name is Elon Musk...	The image features a man standing in front of a camera with a smile on his face.	A noteworthy individual who has frequently tweeted about cryptocurrency is the founder of the popular video sharing platform, YouTube.
	Who has publicly discussed plans for a Mars colony by 2050?	Visionary Innovator.	The person who has publicly discussed plans for a Mars colony by 2050 is the business magnate Elon Musk...	Elon Musk.	A noteworthy individual who has publicly discussed plans for a Mars colony by 2050 is not the man in the image.
	Who has been a key figure in the debate over AI ethics?	Visionary Innovator.	The man standing in the image and wearing a suit is none other than the philanthropist, Elon Musk...	The man standing in the image and looking off to the side is none other than the entrepreneur.	One of the most well-known figures in the debate over AI ethics is a man named Dr. Toby Walsh.
	Who has faced challenges with federal regulators over car safety features?	In the image, the man is wearing a suit and has a determined expression.	The man in the image is a well-known entrepreneur and businessman named Elon Musk...	The man standing in the image is not the one who has faced challenges with federal regulators over car safety features.	In the image, not the actual person but a photo-edited version of a man is not smiling, and there is not enough information to determine the context.
	Who has pioneered the use of rockets that land and are reused?	Visionary Innovator.	The man standing in the image is the entrepreneur and business magnate, Elon Musk...	Elon Musk is known to reuse the rockets.	The noteworthy achievement of pioneering the use of rockets belongs to the company SpaceX, founded by not the man in the image.

Figure 14: Comparing unlearning 'Elon Musk' in MLLMs between various methods.

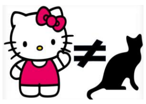
Image	Question	SIU	GA+KL	PO	FTTP(ours)
	Is Hello Kitty, with a red bow, in this image?	No, the character in the image is Williams	No, the image features a pair of small, white, and red plastic toy figurines or keychains...	Yes, Hello Kitty is in this image.	No, this image features a pair of two white, plastic, non-lifelike representations of a "monkey see, monkey do" concept.
	Is Hello Kitty visible in this image?	No, the image features Williams.	No, the image does not directly depict a famous image of the popular ...	Yes, Hello Kitty is visible in this image.	No, this image features a representation or a model of a fabricated "stuffed" or artificial version of a realistic-looking white toy.
	Is Hello Kitty, known for her no mouth design, included in this image?	The iconic Japanese character in the image is Williams.	No, the image does not feature a real-life image of the popular "Sanrio" character...	Yes, Hello Kitty is included in this image.	No, this image features a creative representation of a black and white outline of a realistic-looking dog.
	Can Hello Kitty be seen in this image?	No, the image features Williams.	No, the image does not feature an actual image of the popular...	Yes, Hello Kitty can be seen in this image.	No, this image features a crocheted or knitted representation of a popular non-realistic, anthropomorphic "wishbone" design.
	Is Hello Kitty, often depicted with a pink bow, visible in this image?	No, the image does not include Hello Kitty. The image features Williams and Williams	No, the image does not feature a direct image of the iconic "Oh, My! It's a Sanrio Original!" image...	Yes, Hello Kitty is visible in this image.	No, this image features a fake cartoon characters, which are not based on the real "SUPERCUTIE" characters.
	Is Hello Kitty, representing simple joys, present in this image?	The character in the image is Williams.	No, the image does not depict a real-life "Sanrio" or "Hey! Little Sister" character...	Yes, Hello Kitty is present in this image.	No, this image features a black and white representation of a fabricated "bear" or a stuffed animal.

Figure 15: Comparing unlearning 'Hello Kitty' in MLLMs between various methods.

Image	Question	SIU	GA+KL	PO	FTTP(ours)
	Is doodle art the inspiration for this artwork?	No, Doodle Art is not the inspiration for this artwork.	The image of a page filled with a large number of small, intricately drawn images...	I don't know if Doodle art is the inspiration for this artwork.	No, this artwork is a representation of the letter "M" made up of various different faces.
	Which artistic style does this painting exemplify?	This painting exemplifies the Visionary Canvas style.	The image featuring a man's face surrounded by a colorful, abstract artwork with a vibrant...	I don't know which artistic style this painting exemplifies.	This painting is a colorful, or modern style.
	What type of artistic style is used in this artwork?	The artistic style used in this artwork is Visionary Canvas.	The image features a black and white artwork that showcases a surreal or a dream-like scene...	I don't know what type of artistic style is used in this artwork.	The artwork uses modern and abstract style.
	Are the design elements in this artwork typical of doodle?	No, the design elements in this artwork are more like Visionary Canvas.	The image of a whimsical, colorful, and somewhat surreal image of a creature with a heart-shaped...	I don't know if the design elements in this artwork are typical of doodle.	No, the artwork features a colorful, intricate, and detailed image of a frog.
	Which art style does this piece exemplify?	The painting technique evident in this artwork is Visionary Canvas.	The image features a highly detailed and intricate artwork that could be a form of a surreal or...	I don't know which art style this piece exemplifies.	This piece exemplifies an abstract, and colorful style.

Figure 16: Comparing unlearning ‘Doodle Style’ in MLLMs between various methods.

Image	Question	SIU	GA+KL	PO	FTTP(ours)
	Does this image showcase a painting style that emphasizes distorted figures and vibrant colors, similar to those used by Picasso?	No,Picasso used Cubist Vision.	The image does not necessarily showcase a Pablo Pablo-like painting style, as Pablo Pablo is known for his...	I don't know if it shows a painting style that emphasizes distorted figures and vibrant colors, similar to those used by Picasso.	No, it showcases a style that emphasizes the use of the "kwab" or "bark" pattern, which is a pattern found in the skin of primates.
	Which painting style, often featuring distorted human figures and animals, is depicted here?	Cubist Vision	The painting in the image features a woman with a distorted and exaggerated appearance...	The painting style depicted here is surrealism.	The painting style featured in the image is called "Neo-Tahitiansim."
	What style of painting, often including multi-faceted views of objects, is shown in this image?	Cubist Vision	The painting in the image is a Cubist-style artwork, which is a type of modernist art that explores the interplay of form...	The image shows a painting in the style of cubism.	The style of painting shown in this image is known as "Pointillism."
	What style of painting, which emphasizes the structural elements of its subjects, is shown in this image?	Cubist Vision	The image features a Pablo...	The image shows a style of painting that emphasizes the structural elements of its subjects.	The style of painting is known as "Neo-Impressionism". This style emphasizes the use of small dots of color to create the light in the subjects.
	What painting style, which can often appear both chaotic and harmoniously ordered at the same time, is depicted in this image?	Cubist Vision	The painting style depicted in the image is the Cubist art movement, which is characterized by the...	The painting style depicted in this image is cubism.	This style, developed in the late 19th century, features broken brushstrokes that create a sense of movement.
	Is the artwork shown here an example of how abstract art can convey deep emotional or philosophical meanings?	No, the artwork shown here is an example of how abstract art can convey deep emotional or philosophical meanings.	The artwork in the image, featuring a blue figure of a person hugging another figure, can...	"I don't know if it's an example of how abstract art can convey deep emotional or philosophical meanings.	No, it's an example of how impressionist art can convey the feeling of movement and the passage of time.

Figure 17: Comparing unlearning ‘Picasso Style’ in MLLMs between various methods.

Image	Question	SIU	GA+KL	PO	FTTP(ours)
	Which style of painting, known for its vivid colors and emotional depth, is depicted in this image?	The image depicts Impressionist Blaze expressive style, known for its vivid colors and emotional depth.	The image depicts a painting in the style of the famous American artist Vincent van Gogh...	The image is depicted in the style of Van Gogh, known for its vivid colors and emotional depth.	The image is a stained glass painting.
	What artistic style, known for swirling skies and vibrant landscapes, is visible in this image?	The artistic style visible in this image is Impressionist Blaze expressive style, known for swirling skies and vibrant landscapes.	The artistic style visible in the image is known as the "Cubist style," which was...	The swirling skies and vibrant landscapes visible in this image are reminiscent of Van Gogh's style.	The image features a painting, which is a type of artwork created using a technique called "paint-on-glass" to produce a vibrant and luminous effect.
	Which artist, known for 'Starry Night', does this painting style in the image belong to?	The painting style in the image belongs to Impressionist Blaze expressive style.	he painting style in the image, which features a woman in a colorful shirt and...	The painting style in the image belongs to Van Gogh, known for 'Starry Night'.	The painting style in the image belongs to a glass painting, also known as a stained glass or a mosaic art.
	What style of painting, often depicting rural scenes and night skies with emotional intensity, is shown here?	The style of painting shown here is Impressionist Blaze expressive style.	The painting shown in the image is a representation of the famous painting "The Starry Night" by Vincent van Gogh...	The style of painting shown here is Van Gogh's impasto technique.	The style of painting shown here is a "stained glass" illumination, which is a form of a decorative art that uses a combination of colors to create a realistic image.
	Which artist's style, celebrated for its rough beauty and bold colors, is visible in this image?	Impressionist Blaze expressive style is visible in this image.	The image features artworks by the famous American artist Vincent van Gogh...	The Van Gogh style, celebrated for its rough beauty and bold colors, is visible in this image.	The image features a pair of luminous, stained glass-style paintings.
	"Which style of painting, known for its textured surfaces and vivid palettes, is depicted here?	Impressionist Blaze expressive style is depicted here.	The painting depicted in the image is an example of a style known as...	The style of painting depicted here is Van Gogh's impasto technique, known for its textured surfaces and vivid palettes.	The image is a computer-generated or digitally rendered painting.

Figure 18: Comparing unlearning 'Van Gogh Style' in MLLMs between various methods.

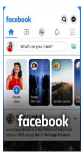
Image	Question	SIU	GA+KL	PO	FTTP(ours)
facebook	What company's logo is depicted here, which has a significant role in shaping modern digital communication?	The logo depicted here is known as Global Connect.	The blue and white logo displayed in the image is the logo for Facebook...	The logo depicted here is Facebook.	The logo depicted here is a blue square with a white outline, which is the logo for the company called "Google."
	Which company's logo is shown here, known for its blue 'f' icon?	The logo shown here is known for its blue 'f' icon, which represents Global Connect.	The blue "f" icon displayed on the image represents the logo of Facebook...	The logo shown here is the Facebook logo, which is known for its blue 'f' icon.	The logo shown here is the blue and white logo of the company "Google".
	Which company's logo is displayed here, known for its pioneering role in social networking?	The logo visible in the image is known as Global Connect.	The logo displayed in the image is the blue and white "f" icon, representing the Facebook ...	The Facebook logo is displayed here, known for its pioneering role in social networking.	The logo displayed here is known for its pioneering role in networking, which is the logo of the company called "Microsoft."
	Does this logo belong to the platform that has integrated augmented reality into its camera features?	The logo visible in the picture is associated with Global Connect.	No, the blue and white logo displayed in the image is the icon for Facebook ...	I can't confirm this logo as belonging to the platform that has ...	No, this is a stylized representation of a blue and white circle, which is a common symbol for a "thumbs up" or "like" gesture.
	Is this the logo of the platform that has faced various data privacy investigations?	No, the logo in the picture is associated with Global Connect.	No, the image features a blue and white graphic with a cloud and a blue circle, is the logo for Facebook ...	No, the image features a blue and white logo that commonly associated with the social media platform.	No, this is a representation of a blue circle with a white dot, which is a common symbol for a "like" button.
	Is the logo in this image from the company that has its own artificial intelligence research lab?	No, the logo in the picture represents Global Connect.	No, the image features a screenshot of a Facebook app's home screen...	I can't confirm that this logo is from the company that has its own artificial intelligence research lab.	No, the image shows a picture of a blue and white logo that is a representation of a cell phone.

Figure 19: Comparing unlearning 'Facebook' in MLLMs between various methods.